Correlation-Based Burstiness for Logo Retrieval

Jerome Revaud
INRIA Grenoble
jerome.revaud@inria.fr

Matthijs Douze
INRIA Grenoble
matthis.douze@inria.fr

Cordelia Schmid
INRIA Grenoble
cordelia.schmid@inria.fr

ABSTRACT

Detecting logos in photos is challenging. A reason is that logos locally resemble patterns frequently seen in random images. We propose to learn a statistical model for the distribution of incorrect detections output by an image matching algorithm. It results in a novel scoring criterion in which the weight of correlated keypoint matches is reduced, penalizing irrelevant logo detections. In experiments on two very different logo retrieval benchmarks, our approach largely improves over the standard matching criterion as well as other state-of-the-art approaches.

Categories and Subject Descriptors: H.2.10 [**Image processing and computer vision**]: Scene analysis

Keywords: Image retrieval; burstiness; correlation.

1. INTRODUCTION

Logo retrieval in real-world images is crucial to quantitatively measure the exposure of brands. In this context, a given logo must be detected and localized in a large collection of images. The preprocessing of the query logo may be costly, as long as the search in the images is fast. This is in contrast with content-based image search engines where the end-to-end search time must be minimized.

Recently, a number of works have tackled logo detection using Bag-of-Words (BoW) techniques [1, 7, 9, 13, 16]. This makes sense since logo images typically contain sharp and contrasted regions that are well handled by keypoint-based representations. Also BoW, in combination with an inverted file, can quickly process large image collections.

The basic BoW model neglects the probabilistic dependence between local features [10]. Yet, our observations show that the independence model does not hold: some pairs, triples or n-uplets of keypoints often appear together in incorrect detections. As illustrated in Figure 1, this is due to the presence of local substructures in logo images that are frequently found in random images. For instance, the Adidas logo is locally similar to printed text patterns such as "ll", "li" (with appropriate fonts) or building windows, see Figure 1. As a result, such patterns are often mistaken for the true logo.

MM'12, October 29–November 2, 2012, Nara, Japan.

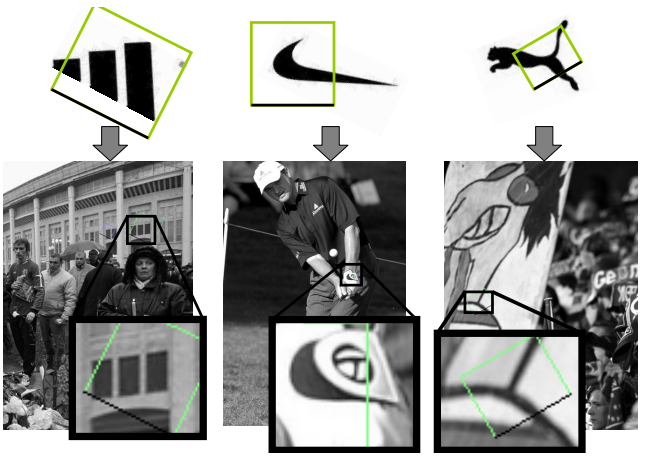

Figure 1: Locally, many logos are similar to patterns frequently found in random photos. The Adidas logo for instance, can partially match with building windows, leading to many detection ambiguities.

In this paper, we show how to down-weight the score of those noisy detections by learning a dedicated burstiness model for the input logo. Section 2 describes how detection hypotheses are obtained by a state-of-the-art approach, then in Section 3 we describe a weighting scheme based on our spatial burstiness approach. Finally, we present experimental results on two logo datasets in Section 4.

2. DETECTING LOGO HYPOTHESES

Given a logo image $\mathcal{L}$, we are interested in detecting all its occurrences in N test images. Some logos might come in different versions, e.g. the Apple logos shown in Figure 4. In this case, we apply the same detection and scoring procedure for each version, and retain the best score across the different versions.

A logo image is represented as a set of K keypoints $\{k_i\}_{i=1..K}$. Because the detection and scoring procedures are identical for each test image, without loss of generality, we consider a single test image with keypoints $\{k'_i\}_{i=1..K'}$.

Keypoint matching. In order to find other instances of the logo, we first compute a set of matches $\mathcal{M} = \{(k_i, k'_j)\}$ between the logo and the image keypoints, based on the quan-

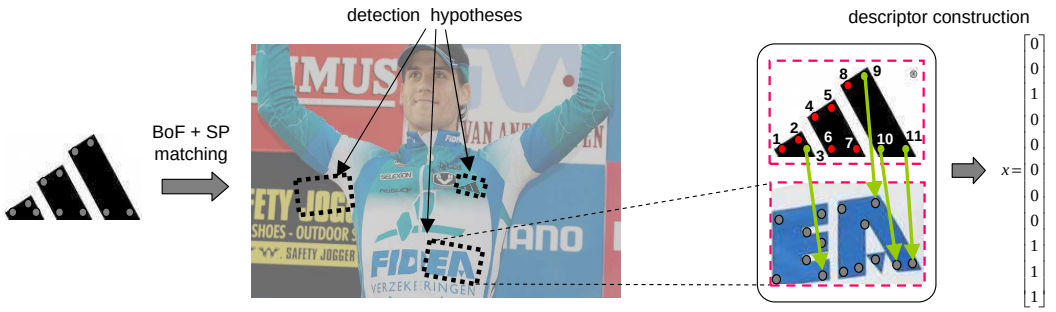

Figure 2: Left: Detecting logo hypotheses. Right: construction of the Boolean descriptor x for a detection hypothesis.

tization of their descriptors into visual words (the codebook size is set to 20,000). To further improve the matching, we compute an additional binary signature for each keypoint and exclude matches for which the Hamming distance between signatures is above 22 [5]. This has been shown to significantly improve over basic quantization at a minor cost.

Spatial verification. Given a set of matches between a logo and an image, we verify the spatial consistency of the matches. As we use scale and rotation invariant keypoints, there exists a single similarity transformation that maps a keypoint to another [14]. We exhaustively examine the similarity transforms corresponding to each match (k, k') in $\mathcal{M}$, as in [11, 14]. For each of those transforms, we count the number of inliers with a two-way transfer error [4]. Because logos are small and usually printed on flat surfaces, a similarity transform is an adequate approximation to the full homographic model and much faster to estimate. We set the threshold on the position error to 5 pixels plus a term that depends on the keypoint's scale. The procedure returns the 4 similarities yielding the maximum number of inliers, with non-maxima suppression of overlapping detections (Figure 2, left).

Descriptor calculation. For each returned hypothesis, we compute a Boolean descriptor $x \in \{0,1\}^K$, where $x_i = 1$ iff. k_i has a matching point k'_j under the transformation hypothesis (Figure 2, right).

3. IMPROVED SCORING OF HYPOTHESES

We now explain how to score each hypothesis according to its descriptor x.

3.1 Baseline

Assuming that all components of x are independent, the optimal score in the Bayesian sense is the ratio of posterior probabilities for the presence of the logo $\mathcal{L}$ and its absence $\neg\mathcal{L}$:

$$\frac{p(\mathcal{L}|x)}{p(\neg\mathcal{L}|x)} \propto \prod_i \frac{p(x_i|\mathcal{L})}{p(x_i|\neg\mathcal{L})}. \quad (1)$$

Assuming uniform distributions for $p(x_i|\mathcal{L})$ and $p(x_i|\neg\mathcal{L})$, and taking the logarithm, we obtain the number of inliers:

$$s_{\text{bsl}}(x) = \log \frac{p(\mathcal{L}|x)}{p(\neg\mathcal{L}|x)} \propto \sum_i x_i. \quad (2)$$

This scoring criterion is widely used in RANSAC implementations [2, 7, 15].

Figure 3: Groups of keypoints (shown as ellipses) obtained after correlation-based clustering (section 3.3).

3.2 Burstiness handling

In image retrieval, it has been observed that the independence assumption on $p(x_i|\mathcal{L})$ and $p(x_i|\neg\mathcal{L})$ does not hold. Visual elements can appear more frequently in an image than a statistically independent model would predict. This phenomenon, initially observed in text retrieval, is called *burstiness* [6, 8, 12]. Jégou et al. [6] use the following improved score in image retrieval. They group matches with identical visual word indexes and apply a down-weighting scheme (i.e. square-rooting) on the score of each group:

$$s_{\text{burst}}(x) = \sum_{g=1}^{h} \sqrt{\sum_i x_i G_{i,g}}. \quad (3)$$

$G \in \{0,1\}^{K\times h}$ is the "grouping matrix": $G_{i,j} = 1$ if k_i is quantized into visual word j (G's rows sum to one)[1]. This score works well in practice because square-rooting approximates a probabilistic model of the visual word dependencies [3].

3.3 Proposed weighting

Instead of grouping keypoints according to their visual word index, we propose to learn a new grouping matrix H that better describes the burstiness in logo detection. The intuition is to group keypoints frequently matched together in incorrect detections (Figure 3), so as to down-weight their score as a whole. To learn their statistical distribution, we first measure the correlations between keypoints matched in incorrect hypotheses. Then, clustering correlated matches yields H.

Let $X = [x_1 \dots x_L] \in \{0,1\}^{K\times L}$ the matrix of L descriptors collected on a training dataset disjoint from the test set, from which the logo is known to be absent. All these descriptors correspond to incorrect hypotheses (Figure 1). Let $C \in \mathbb{R}_+^{K\times K}$ be the corresponding correlation matrix[2],

[1] In [6], the formula is expressed differently but it is equivalent for binary scores.

[2] Strictly speaking, the columns of X should be centered, but the mean of each column is approximately 0, so it makes little difference.

with cells defined as:

$$C_{ij} = \frac{\mathrm{cov}(X_i, X_j)}{\sigma_{X_i}\sigma_{X_j}} = \frac{\sum_\ell x_{i\ell}x_{j\ell}}{\sqrt{\left(\sum_\ell x_{i\ell}^2\right)\left(\sum_\ell x_{j\ell}^2\right)}} \quad (4)$$

Elements of C range in $[0,1]$ and represent how frequently two keypoints jointly appears in X. To extract groups, we perform hierarchical clustering [17] in the space of logo keypoints using C as similarity metric, see Figure 3 for example clusters. This yields the grouping matrix H. In our experiments, we have set the minimum linkage threshold [17] to 0 (yielding the minimum number of clusters), which produced the best results. Hereafter, we also investigate more general weighting schemes and soft-assigned clustering.

Generalized burstiness. We generalize the burstiness formulation of eq. (3) into

$$s_{\mathrm{corr}}(x) = \sum_{g=1}^{h} \omega\left([x_1 H_{1,g}, \ldots, x_K H_{K,g}]\right), \quad (5)$$

where $\omega : \{0,1\}^K \to \mathbb{R}^+$ is down-weighting function. Among many possibilities, we experiment square-rooting $\omega(x) = \sqrt{\sum_i x_i}$ as in eq. (3), $\omega(x) = \log(1+\sum_i x_i)$, and the family of norms

$$\omega(x) = \|x\|_\gamma \quad (6)$$

with $\gamma > 1$. All those functions obey the desirable property that

$$\forall a_1 \ldots a_n \in \{0,1\}^n, \quad \sum_i \omega(a_i) \geq \omega([a_1 \ldots a_n]), \quad (7)$$

which guarantees that a small number of points spread into a large number of groups score higher than the same number of points belonging to a single group.

Soft-assigned groups. Because soft-assignment has been shown to be beneficial for image retrieval [14], we evaluate a more flexible grouping procedure. We extend the Boolean matrix H to a real-valued matrix $S \in [0,1]^{K\times h}$. The scoring function (5) generalizes smoothly if we choose S with rows summing to one. Also, we simply assume that the number of groups h is equal to K: there is one "soft-group" per logo keypoint k_i.

A natural choice for the matrix S is to build it directly from the correlation matrix C. We use the following formulation:

$$S_{i,j} = \frac{C_{i,j}^\alpha}{\sum_j C_{i,j}^\alpha}. \quad (8)$$

The normalization is required so that the rows sum to 1. We raise the elements of C to a power $0 < \alpha \leq 1$, to compensate for the fact that the correlation values are underestimated due to the asymmetric repeatability noise in keypoint extraction and matching. Typically, $\alpha = 0.4$ yields good results (see Section 4).

In the experiments, we evaluate both the hard and soft-assigned grouping H and S, respectively, in addition to different down-weighting schemes ω.

4. EXPERIMENTS

In this section, we present experimental results on two different datasets introduced to evaluate logo retrieval in real-world images: the BelgaLogos [7] and the FlickrLogos [16] datasets.

Table 1: Comparison of the different down-weighting schemes, for hard and soft assignment.

		$\omega(x)$	Belga Qset1	Belga Qset2	Flickr
		None	0.328	0.383	0.577
Hard H	#1	$\sqrt{\sum x_i}$	0.387	0.440	0.658
	#2	$\log(1+\sum x_i)$	0.385	0.440	0.656
	#3	$\|x\|_\gamma,\ \gamma=\infty$	0.388	0.428	0.724
Soft S	#1	$\sqrt{\sum x_i}$	left-out because (7) does not hold		
	#2	$\log(1+\sum x_i)$	0.333	0.389	0.566
	#3	$\|x\|_\gamma,\ \gamma=\infty$	**0.414**	**0.481**	**0.726**

Figure 4: Sample logos from BelgaLogos (set #2) [7] (top row) and from Flickr-Logos [16] (bottom row).

4.1 BelgaLogos dataset

The BelgaLogos dataset [7] was created in collaboration with the Belgavox press agency. It is composed of 10,000 press photographs annotated for 26 logos. The set Qset1 is composed of 55 queries, each defined by an image from the database and the logo's bounding box in this image. The set Qset2 is composed of 26 thumbnails, representing the "ideal" logos, see top row in Figure 4.

As in [7], we use SIFT keypoints [11] and report mean Average-Precision (mAP) results for both sets. We introduce several optional improvements: DBL – doubling the image size prior to keypoint extraction, because the logo often appears at small scales in database images; INV – generating color-inverted versions of monochromatic logos like Adidas or Nike, that can be printed either black-on-white or white-on-black; MA – assigning multiple visual words to each query keypoint [5].

For our approach, we estimate the correlation matrix C from an external dataset composed of 15,000 images collected from Flickr. We manually verified that no logo appears in them, using the baseline method and checking the 10 top retrievals for each logo.

4.2 FlickrLogos dataset

Romberg et al. [16] built the FlickrLogos dataset by downloading real-world images including one of 32 logos from Flickr. The dataset is very different from BelgaLogos, because logos are more textured, usually large, and come in several versions. Sample images are shown in Figure 4. The dataset is partitioned into 3 subsets: P_1 contains 10 images per logo, chosen to contain little clutter and noise. Each of P_2 and P_3 contains 30 images per logo and 3,000 additional background images where no logo appears. Following the protocol of [16], P_1 serves as query set and results are reported in terms of average recall on P_3, while P_2 is used as training set to learn per-logo recognition thresholds (the target precision is set to 95% as in [16]).

In order to estimate C and learn the recognition thresh-

Table 2: mAP performance for various options on the BelgaLogos dataset (Section 4.1).

options	Qset1			Qset2		
	s_{bsl}	s_{burst}	s_{corr}	s_{bsl}	s_{burst}	s_{corr}
default	0.205	0.216	0.250	0.182	0.215	0.250
+DBL	0.286	0.302	0.367	0.206	0.226	0.283
+INV	-		-	0.300	0.329	0.388
+MA	0.328	0.344	**0.414**	0.383	0.404	**0.481**
state of the art [7]	0.341			0.257		

Table 3: Recall performance on the FlickrLogos dataset (precision not shown but always $\geq$ 98%).

	s_{bsl}	s_{burst}	s_{corr}
proposed method	0.577	0.614	**0.726**
state of the art [16]	0.61		

olds at the same time, we perform 10-fold cross-validation on P_2.

4.3 Impact of the parameters

Table 1 compares different down-weighting schemes ω that were proposed in [6] (#1, #2) and in this paper (#3), eq. (6). The proposed scheme (#3) depends upon parameters γ, plus α in the case of soft-assignment (eq. (8)). To evaluate their impact, we fix in turn γ and α to their optimal values and vary the other parameter. Overall, all examined values of γ and α yield good performance, with better results for larger values of γ and for $\alpha \in [0.3, 0.5]$. In the following, we thus set $\gamma = \infty$ (i.e. $\omega(x) = \|x\|_\infty = \max(x)$) and $\alpha = 0.4$.

In Table 1, we evaluate different weighting schemes ω for our scoring method s_{corr}. The proposed weighting (#3) yields excellent performance: it counts the number of detected groups rather than the number of inlier points. In particular, it clearly outperforms the other ones for all datasets when combined with soft-assignment. It is also worth noting that the logarithmic down-weighting scheme (#2) only performs well for the case of hard-assignment. In the following, we thus use $\omega(x) = \|x\|_\infty$ and soft-assignment.

4.4 Quantitative results and conclusion

Table 2 shows quantitative results for the BelgaLogos dataset. The proposed scoring constantly improves over the baseline or the classical burstiness model of [6]. The latter approach improves slightly by about 2%. In contrast, the proposed learned burstiness model outperforms the baseline by up to 9% and 10% on Qset1 and Qset2 for the best setting, i.e. with DBL, INV and MA. Due to the specificity of the BelgaLogos dataset, in which logos often appears at very small scales and with inverted colors (Figure 5), DBL, INV and MA are also important to obtain good performance. On FlickrLogos (Table 3), the proposed method largely improves over the baseline (+15%) and over the classical burstiness model [6] (+3.7%).

Compared to the state of the art, our approach outperforms the a-contrario query expansion of Joly et al. [7] by 7% on Qset1 and 22% on Qset2. In contrast to this method, our approach does not add any overhead at run-time and could, in fact, be combined with query expansion. On the FlickerLogos dataset, we outperform the method of [16] by 12%. Contrary to the cascaded triplets features of [16], our approach remains generic and can be added to existing BoW-based systems.

Figure 5: Examples of detections output by our system for the 3 queries on the left.

Acknowledgments

This work was realized as part of the Quaero Project, funded by OSEO, French State agency for innovation.

5. REFERENCES

[1] Andrew D. Bagdanov, Lamberto Ballan, Marco Bertini, and Alberto Del Bimbo. Trademark matching and retrieval in sports video databases. In *MIR*, 2007.

[2] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *ICCV*, 2007.

[3] Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid. Image categorization using non-iid image models. In *CVPR*, 2012.

[4] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004.

[5] Hervé Jégou, Matthijs Douze, and Cordelia Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *ECCV*, 2008.

[6] Hervé Jégou, Matthijs Douze, and Cordelia Schmid. On the burstiness of visual elements. In *CVPR*, 2009.

[7] Alexis Joly and Olivier Buisson. Logo retrieval with a contrario visual query expansion. In *ACM Multimedia*, 2009.

[8] Slava M. Katz. Distribution of content words and phrases in text and language modelling. *Natural Language Engineering*, 1996.

[9] Jim Kleban, Xing Xie, and Wei-Ying Ma. Spatial pyramid mining for logo detection in natural scenes. In *ICME*, 2008.

[10] David D. Lewis. Naive (bayes) at forty: The independence assumption in information retrieval. In *ECML*, 1998.

[11] David G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60:91–110, 2004.

[12] Rasmus E. Madsen, David Kauchak, and Charles Elkan. Modeling word burstiness using the Dirichlet distribution. In *ICML*, 2005.

[13] Jingjing Meng, Junsong Yuan, Yuning Jiang, Nitya Narasimhan, Venu Vasudevan, and Ying Wu. Interactive visual object search through mutual information maximization. In *ACM Multimedia*, 2010.

[14] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *CVPR*, 2008.

[15] J. Philbin, M. Isard, J. Sivic, and A. Zisserman. Descriptor learning for efficient retrieval. In *ECCV*, 2010.

[16] Stefan Romberg, Lluis Garcia Pueyo, Rainer Lienhart, and Roelof van Zwol. Scalable logo recognition in real-world images. In *ICMR*, 2011.

[17] R. Sibson. SLINK: An optimally efficient algorithm for the single-link cluster method. *The Computer Journal*, 16:30–34, 1973.